Introduction to Bayesian inference

End of course data analysis project

—

The Center for Disease Control (CDC) reports the vaccination coverage of Varicella among young children. Varicella, commonly known as chicken- pox, is a highly contagious viral infection caused by the varicella-zoster virus (VZV). Vaccination against chickenpox has been highly effective in reducing the incidence and severity of the disease. In the United States, vaccination against varicella has been part of the routine childhood immunization sched- ule since the mid-1990s. Since the vaccine's introduction, there has been a dramatic decline in the number of chickenpox cases, hospitalizations, and deaths associated with the disease. The target for vaccination coverage of varicella (chickenpox) in the United States, as set by the Centers for Dis- ease Control and Prevention (CDC), is typically around 90% or higher for children. This high coverage rate is aimed at achieving herd immunity and preventing outbreaks of chickenpox within communities.

Project 1: Insurance

The next table summarizes, based on a survey, the number of children in the birth cohort 2014-2017 that had at least one dose of the Varicella vaccine. It gives the number of vaccinated children (Vaccinated) amongst the number of children in the survey (Sample Size). The information is provided for 5 regions of the US, and split according to insurance status (private insurance, uninsured or any Medicaid).

Geography	Insurance	Vaccinated	Sample Size
North Carolina	Any Medicaid	380	419
North Carolina	Private Insurance Only	632	673
North Carolina	Uninsured	28	34
Georgia	Any Medicaid	363	396
Georgia	Private Insurance Only	527	576
Georgia	Uninsured	36	50
Wisconsin	Any Medicaid	282	332
Wisconsin	Private Insurance Only	514	548
Wisconsin	Uninsured	16	34
Florida	Any Medicaid	446	490
Florida	Private Insurance Only	588	628
Florida	Uninsured	28	39
Mississippi	Private Insurance Only	400	441
Mississippi	Uninsured	27	32

Question 1

> Derive analytically the posterior of the vaccination coverage per ge- ography and insurance group. Use a conjugate prior that (1) reflects no knowledge on the vaccination coverage, and (2) reflects that vac- cination coverage is typically around 90% or higher. Give posterior summary measures of the vaccination coverage per geography and in- surance group. Is the choice of the prior impacting your results?

Theoretical considerations

The outcome *Vaccinated/Not Vaccinated* follows a Bernouilli distribution with parameter p:

V : Vaccination status $V \in \{0, 1\}$ $V \sim \mathcal{Bern}(p)$

It is known from theory that the sum of n $i.i.d$ Bernoulli random variables follows a Binomial distribution. This will be used to model the sample outcome: the number of vaccinated people V_s in a random sample of size n:

$$V_s = \sum_i^n V_i \sim \mathcal{Binom}(n, \theta)$$

where θ is the parameter of interest - the vaccine coverage.

In the course, we saw that the Beta distribution is the conjugate prior for binomially distributed data:

Distribution	Formula
Prior	$p(\theta) = \mathcal{Beta}(\alpha, \beta)$
Likelihood	$p(y \mid \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$
Posterior	$p(\theta \mid y) = \mathcal{Beta}(\alpha + k, \beta + n - k)$

The summary measures for the Beta distribution are defined as follows:

Summary Measure	Formula
Mean	$\frac{\alpha}{\alpha+\beta}$
Median	See Note
Mode	$\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$

Note: The median of the Beta distribution does not have a simple closed form expression. It can be approximated numerically or using statistical software.

```
source("./code.R")
```

```
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 14
   Unobserved stochastic nodes: 3
   Total graph size: 86

Initializing model
```

Choice of prior distributions

(1) No prior knowledge

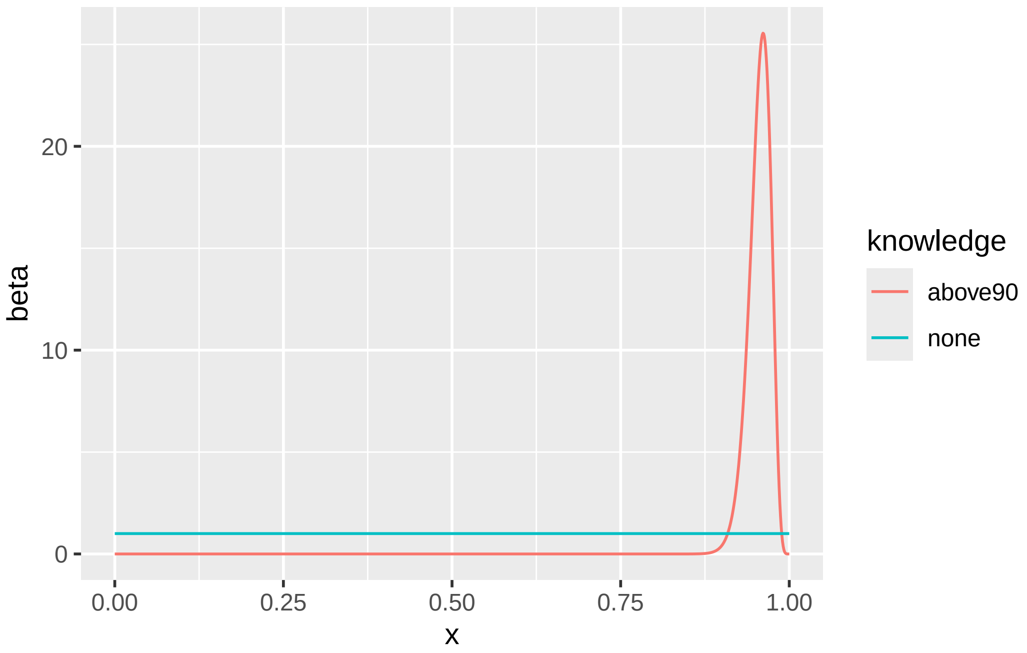
In order to reflect no prior knowledge on the vaccine coverage, the weakly-informative prior Beta(1,1) will be used, which is equivalent to the uniform distribution over $[0, 1]$.

(2) Vaccine coverage >90%

For modeling prior knowledge that vaccine coverage is about 90%, we chose the Beta(150, 7) distribution.

Comparison of priors

```
plot_priors
```



Results

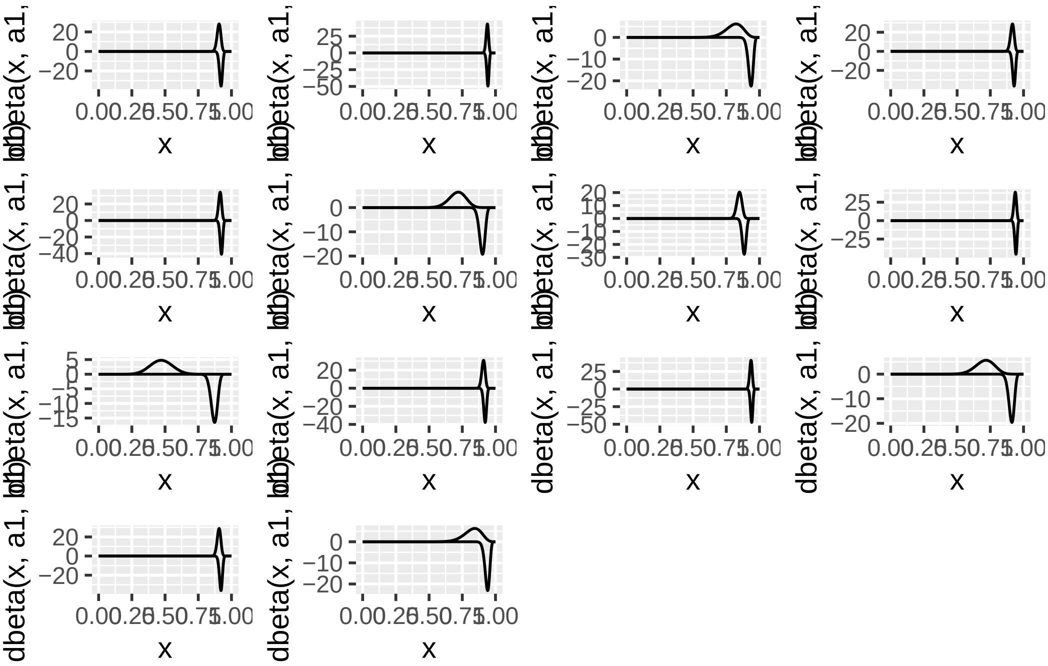
::: .cell ::: .cell-output-display

Table 4: Posterior distribution parameters and summary measures per geography for the two different Beta priors

		Posterior parameters and summary measures													
		Beta(1,1) prior							Beta(150,7) prior						
Geography	Insurance	alpha	beta	mean	mode	var	HPD LL	HPD UL	alpha	beta	mean	mode	var	HPD LL	HPD UL
North Carolina	Any Medicaid	381	40	0.90	0.91	36.29	0.88	0.93	530	46	0.92	0.92	42.40	0.90	0.94
North Carolina	Private Insurance Only	633	42	0.94	0.94	39.45	0.92	0.95	782	48	0.94	0.94	45.28	0.93	0.96
North Carolina	Uninsured	29	7	0.81	0.82	5.80	0.66	0.92	178	13	0.93	0.94	12.18	0.89	0.96
Georgia	Any Medicaid	364	34	0.91	0.92	31.17	0.89	0.94	513	40	0.93	0.93	37.17	0.90	0.95
Georgia	Private Insurance Only	528	50	0.91	0.91	45.75	0.89	0.93	677	56	0.92	0.92	51.79	0.90	0.94
Georgia	Uninsured	37	15	0.71	0.72	10.88	0.58	0.83	186	21	0.90	0.90	18.96	0.85	0.94
Wisconsin	Any Medicaid	283	51	0.85	0.85	43.34	0.81	0.88	432	57	0.88	0.89	50.46	0.85	0.91
Wisconsin	Private Insurance Only	515	35	0.94	0.94	32.83	0.91	0.96	664	41	0.94	0.94	38.67	0.92	0.96
Wisconsin	Uninsured	17	19	0.47	0.47	9.22	0.31	0.63	166	25	0.87	0.87	21.84	0.82	0.91
Florida	Any Medicaid	447	45	0.91	0.91	40.97	0.88	0.93	596	51	0.92	0.92	47.05	0.90	0.94
Florida	Private Insurance Only	589	41	0.93	0.94	38.39	0.91	0.95	738	47	0.94	0.94	44.24	0.92	0.96
Florida	Uninsured	29	12	0.71	0.72	8.69	0.56	0.83	178	18	0.91	0.91	16.43	0.86	0.94
Mississippi	Private Insurance Only	401	42	0.91	0.91	38.10	0.88	0.93	550	48	0.92	0.92	44.22	0.90	0.94
Mississippi	Uninsured	28	6	0.82	0.84	5.09	0.68	0.93	177	12	0.94	0.94	11.30	0.90	0.97

::: :::

```r
grid.arrange(grobs = plots)
```



Question 2

> nvestigate whether the vaccination coverage is associated with the in- surance status using a logistic regression model Yij Binom(ij ,Nij) with logit(ij) = 0 + 1IAnyMedicaid,ij + 2IUninsured,ij where i is the location, j is the insurance status, ij is the vaccination coverage and I. are dummy variables. Assume non-informative priors for the parameters to be estimated. Write and explain the code in BUGS language

https://github.com/andrewcparnell/jags_examples/blob/master/R%20Code/jags_logistic_regression.R

```
model
{
  for (t in 1:T) {
    # Likelihood
    y[t] ~ dbin(p[t], K[t])
    # conditional mean model using link function
    logit(p[t]) <- alpha_0 + ins_1 * x_1[t] + ins_2 * x_2[t]
 }

  # Priors
  alpha_0 ~ dnorm(0.0,0.01)
  ins_1 ~ dnorm(0.0,0.01)
  ins_2 ~ dnorm(0.0,0.01)


  # Vaccine coverage per Insurance group
  pi_private <- exp(alpha_0)/(1+exp(alpha_0))
  pi_medicaid <- exp(alpha_0 + ins_1)/(1+exp(alpha_0 + ins_1))
  pi_uninsured <- exp(alpha_0 + ins_2)/(1+exp(alpha_0 + ins_2))

  # Difference between vaccine coverage per insurance group
  diff_priv_medicaid  <- pi_private - pi_medicaid
  diff_priv_uninsured <- pi_private - pi_uninsured
}
```

Likelihood function and model specification

For each row `t` in the dataset, the outcome `y[t]` (number of vaccinated children in the sample) is specified as drawn from a Binomial distribution with parameters `p[t]` and sample size `K[t]`. We also specify the conditional mean model for this likelihood function using the `logit` link function.

Prior distribution

As prior distribution for all parameters (intercept and indicator variables for the insurance group), a vague prior is chosen : $\mathcal{N}(\mu = 0, \tau = 0.01)$.

Quantities of interest During each MCMC run, the vaccine coverage per insurance group is calculated using the inverse logit transformation, this will give access to the posterior distribution of vaccine coverage per insurance group. Furthermore, the differences between vaccine coverages are calculated which will be needed to answer Question 6.

Question 3

> Run the MCMC method and check convergence of the MCMC chains. Give the details on how you checked convergence.

MCMC run summary

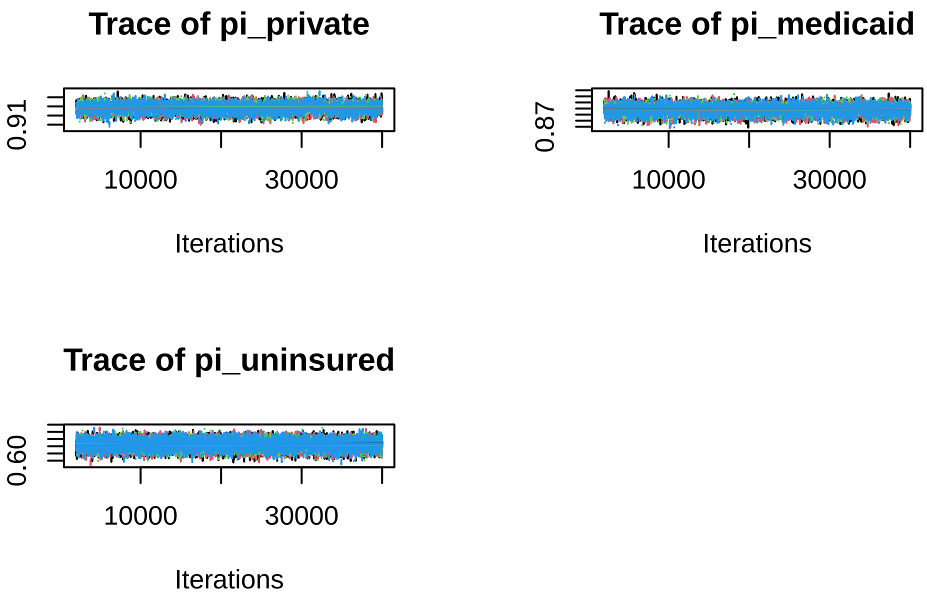
```
model_run
```

```
Inference for Bugs model at "4", fit using jags,
 4 chains, each with 40000 iterations (first 2000 discarded), n.thin = 2
 n.sims = 76000 iterations saved. Running time = 3.333 secs
                    mu.vect sd.vect   2.5%     25%     50%     75%   97.5%
diff_priv_medicaid    0.030   0.009  0.013   0.024   0.030   0.036   0.048
diff_priv_uninsured   0.215   0.033  0.152   0.192   0.214   0.237   0.282
pi_medicaid           0.898   0.007  0.884   0.894   0.899   0.904   0.912
pi_private            0.928   0.005  0.919   0.925   0.929   0.932   0.938
pi_uninsured          0.713   0.033  0.647   0.692   0.714   0.736   0.776
deviance            102.252   2.430 99.480 100.467 101.625 103.365 108.493
                     Rhat n.eff
diff_priv_medicaid  1.002  4100
diff_priv_uninsured 1.001 11000
pi_medicaid         1.001  5500
pi_private          1.003  1500
pi_uninsured        1.001 21000
deviance            1.001  5700

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule: pV = var(deviance)/2)
pV = 2.9 and DIC = 105.2
DIC is an estimate of expected predictive error (lower deviance is better).
```

Convergence checks

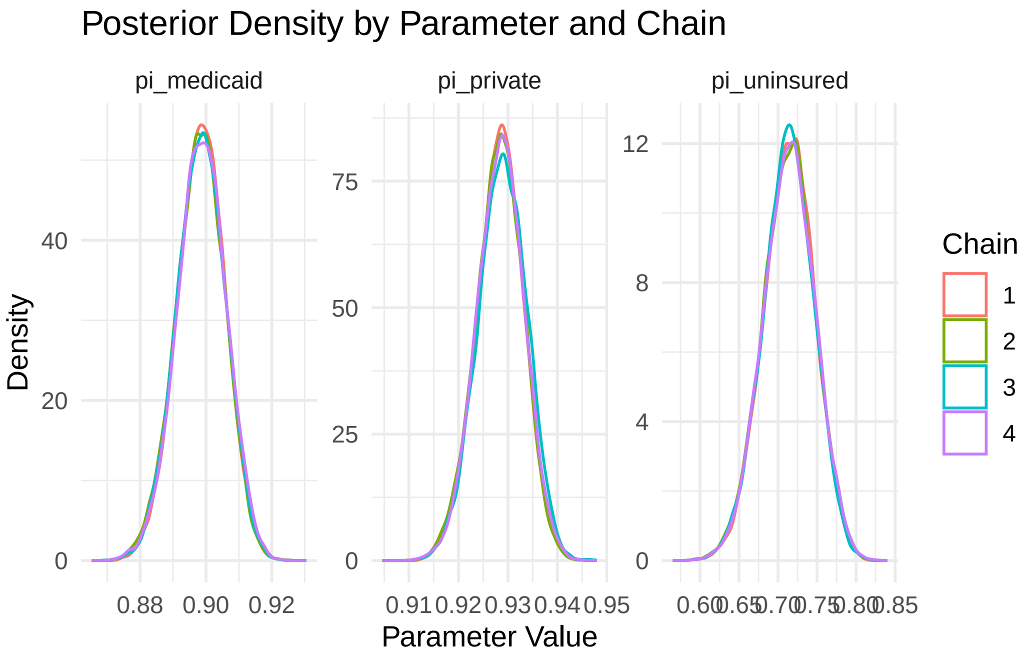
Traceplots

```r
plot(mcmc[,c("pi_private", "pi_medicaid", "pi_uninsured")], trace = T, density = F)
```



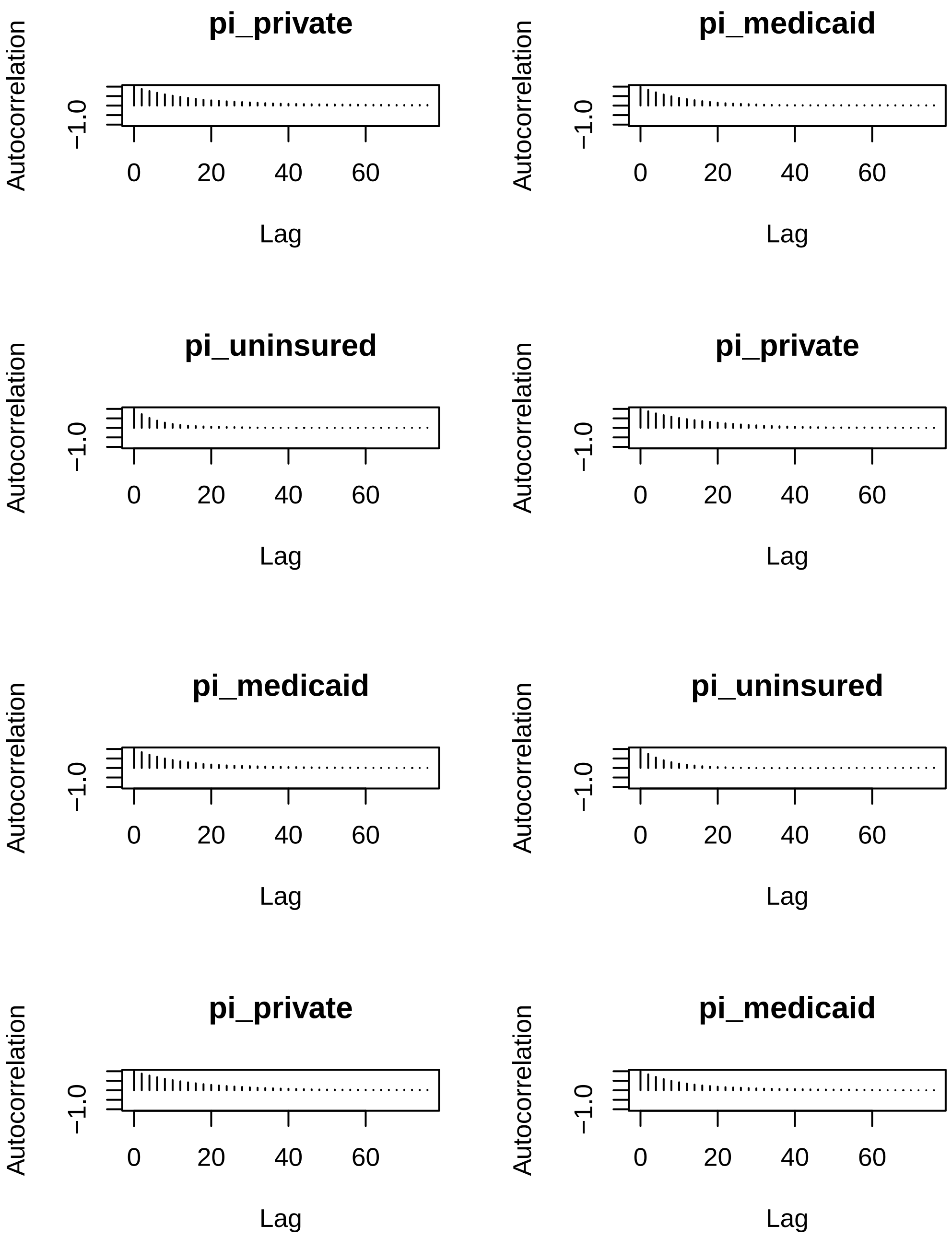
Density plots per chain

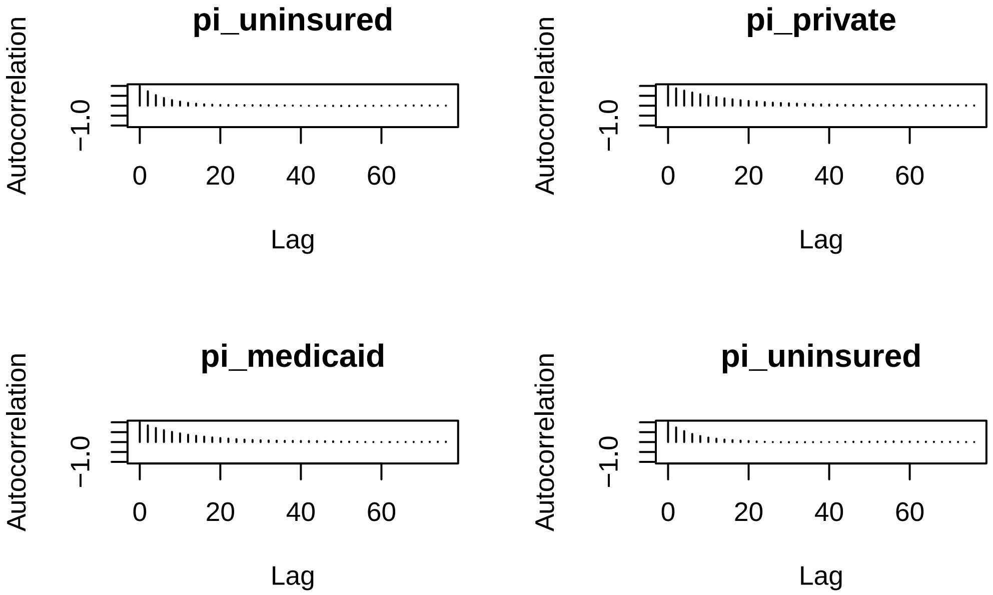
```r
ggplot(dens_plot_df, aes(x = value, color = factor(chain))) +
  geom_density() +
  labs(x = "Parameter Value", y = "Density",
       title = "Posterior Density by Parameter and Chain",
       color = "Chain") +
  facet_wrap(~ parameter, scales = "free") + # Use facet_wrap on the 'parameter' column
  theme_minimal()
```



Autocorrelation plot

```r
autocorr.plot(mcmc[,pi_params])
```


pi_private
pi_medicaid
pi_uninsured
pi_private
pi_medicaid
pi_uninsured
pi_private
pi_medicaid
Autocorrelation
−1.0
0
20
40
60
Lag

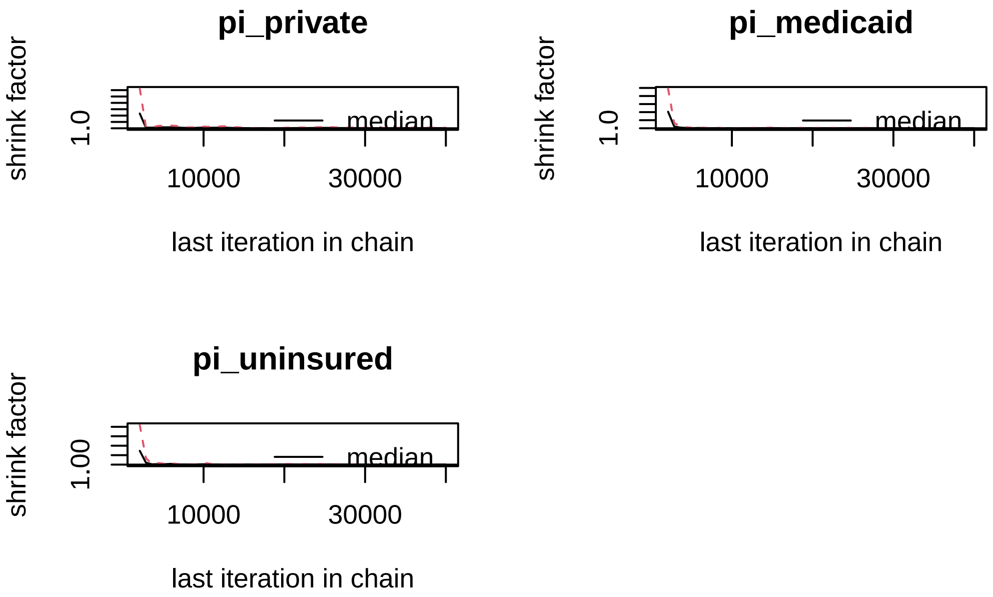


$\hat{R}$

```
gelman.diag(mcmc[,pi_params])
```

```
Potential scale reduction factors:

             Point est. Upper C.I.
pi_private            1       1.01
pi_medicaid           1       1.00
pi_uninsured          1       1.00

Multivariate psrf

1
```

```
gelman.plot(mcmc[,pi_params])
```

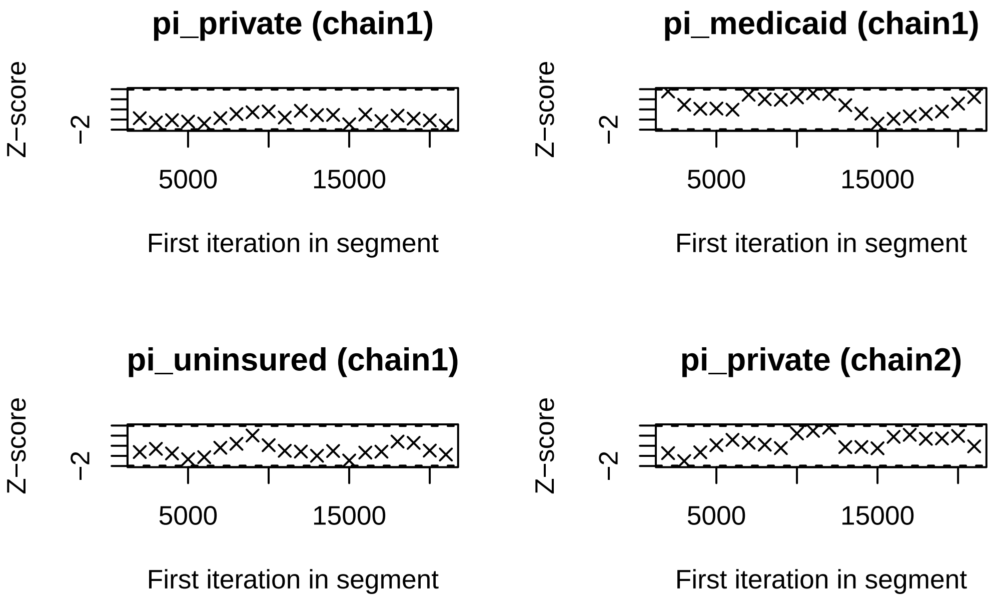


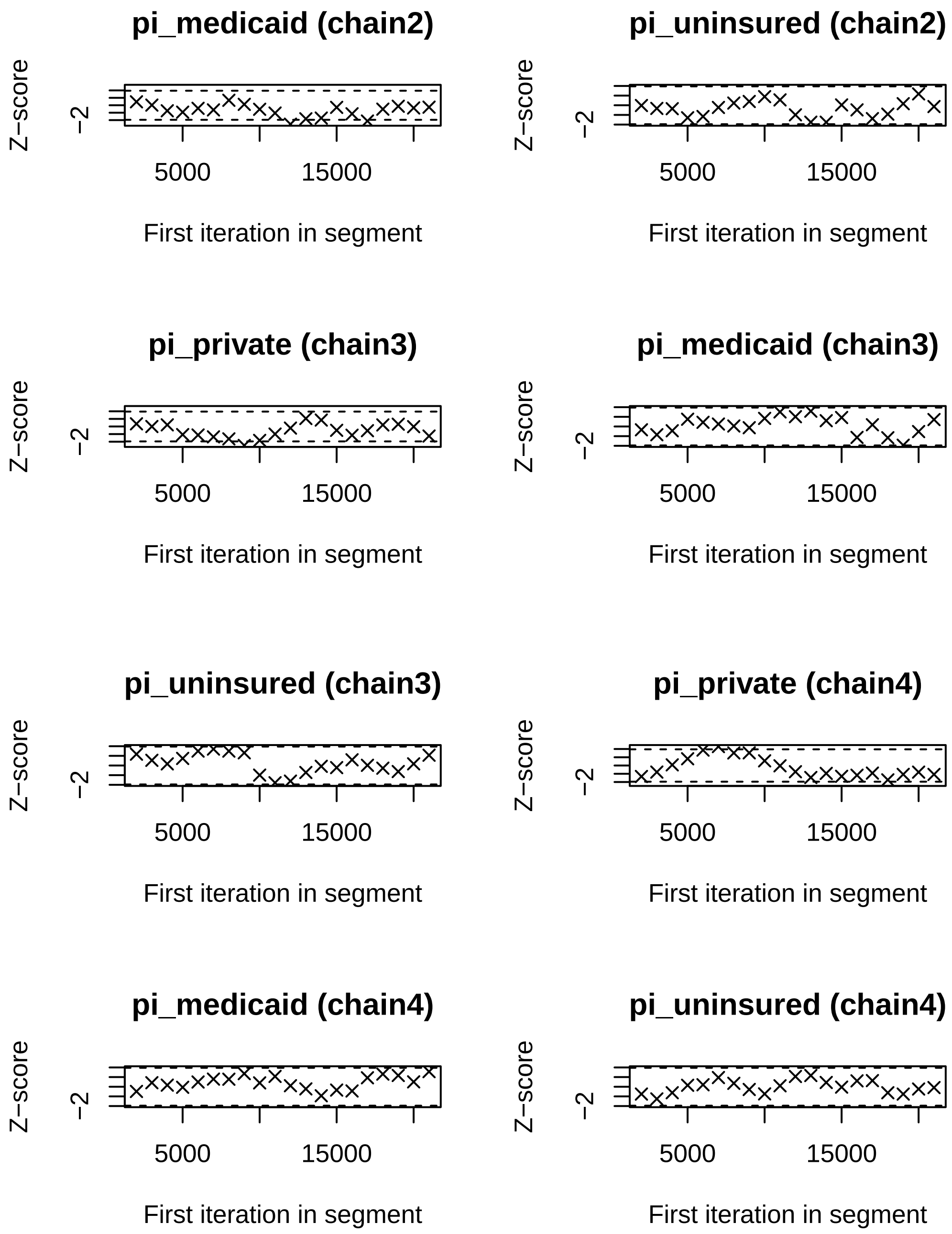
Geweke diagnostics

```
geweke.diag(mcmc[,pi_params])
```

```
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

 pi_private  pi_medicaid pi_uninsured
    -0.8647       1.7742      -0.6152


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

 pi_private  pi_medicaid pi_uninsured
   -0.72767      0.46469     -0.02443


[[3]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

 pi_private  pi_medicaid pi_uninsured
     0.3522      -0.3420       1.1743


[[4]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

 pi_private  pi_medicaid pi_uninsured
    -1.3319      -0.4857      -0.7577
```

```
geweke.plot(mcmc[,pi_params])
```





Question 4

> Make a plot of the posterior of the model parameters and give posterior summary measures. Interpret the results.

Question 5

> Give the posterior estimate of the vaccination coverage per region and insurance status. Compare with the analytical results you obtained in Question 1.

Question 6

> Based on the logistic regression model, what is the probability (a posteriori) that coverage amongst children that have private insurance is higher than amongst children that have any medicaid? And compared to children with no insurance?